A general approach to account for dependence in large-scale multiple testing

Chloé Friguet

Journal de la Société Française de Statistique
Vol. 153 No. 2 (2012)

A general approach to account for dependence in large-scale multiple testing

Titre: Un cadre global pour la prise en compte de la dépendance dans les procédures de tests multiples en grande dimension

Chloé Friguet[1]

Abstract: The data generated by high-throughput biotechnologies are characterized by their high-dimension and heterogeneity. Usual, tried and tested inference approaches are questioned in the statistical analysis of such data. Motivated by issues raised by the analysis of gene expressions data, I focus on the impact of dependence on the properties of multiple testing procedures in high-dimension. This article aims at presenting the main results: after introducing the issues brought by dependence among variables, the impact of dependence on the error rates and on the procedures developed to control them is more particularly studied. It results in the description of an innovative methodology based on a factor structure to model the data heterogeneity, which provides a general framework to deal with dependence in multiple testing. The proposed framework leads to less variability for error rates and consequently shows large improvements of power and stability of simultaneous inference with respect to existing multiple testing procedures. Besides, the model parameters estimation in a high-dimensional setting and the determination of the number of factors to be considered in the model are evoked. These results are then illustrated by real data from microarray experiments analyzed using the R package called FAMT.

This paper is an extended written version of my oral presentation on the same topic at the *44th Journées de Statistique* organized by the French Statistical Society (SFdS) in Bruxelles, Belgium, 2012, when being awarded the Marie-Jeanne Laurent-Duhamel prize.

Résumé : Les données générées par les biotechnologies haut-débit sont caractérisées par leur grande dimension et leur hétérogénéité. L'analyse statistique de ces données remet en cause y compris les approches les plus éprouvées, comme les méthodes usuelles d'inférence statistique. Cet article a pour objectif de présenter une étude de l'impact de la dépendance sur les propriétés des procédures de tests multiples en grande dimension : après une description introductive des principales problématiques liées à la présence de dépendance, les mesures de risques d'erreurs et les algorithmes permettant de contrôler ces risques lors de la mise en œuvre de procédures de tests multiples sont plus particulièrement étudiés. Cette étude analytique aboutit à la définition d'un cadre général de la prise en compte de l'hétérogénéité des données, grâce à la modélisation de la structure de dépendance par Analyse en Facteurs. L'instabilité des procédures induite par la présence de dépendance est alors réduite, procurant à la fois une augmentation de la puissance des tests et une diminution de la variabilité des taux d'erreurs. La mise en œuvre de cette méthode est également évoquée, et les résultats méthodologiques sont illustrés à partir de données génomiques, analysées à l'aide du package FAMT du logiciel libre R qui implémente les méthodes présentées précédemment.

Cet article accompagne la conférence que j'ai eu l'honneur de donner lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des *44èmes Journées de Statistique* organisées par la Société Française de Statistique à Bruxelles, en mai 2012.

Keywords: Multiple testing, Dependence, High-dimension, Error rates, Factor Analysis, Proportion of null hypotheses
Mots-clés : Tests multiples, Dépendance, Grande dimension, Taux d'erreurs, Analyse en facteurs, Proportion d'hypothèses nulles

[1] Lab. de Math. de Bretagne Atlantique (LMBA) - CNRS UMR 6205
Univ. de Bretagne-Sud, Campus de Tohannic - Vannes
E-mail: chloe.friguet@univ-ubs.fr

 ISSN: 2102-6238

1. Introduction: large-scale multiple testing

1.1. Background

Multiple testing The decision of a statistical test requires to choose between two hypotheses: the null hypothesis (H_0) and the alternative one (H_1). The goal of a test procedure is to control the risk of wrongly reject H_0 (type-I error).
Multiple testing refers to the simultaneous tests of several hypotheses. Extending the well-assessed theory of hypothesis testing, issues raised by multiple testing have been widely discussed in the statistical literature since the 1930's when Fisher firstly proposed procedures to test several linear contrasts in analysis of variance. Multiplicity has been an abounding issue and many methods to deal with the number of simultaneous tests are available. Basically, multiple testing procedures rely first on the computation of a p-value for each response variable and then on the choice of a threshold t on the p-values associated to the individual tests. The decision rule states that null hypotheses associated to p-values lesser than t are rejected. Differences between the procedures are due to the way of finding the threshold.
In the first step, the choice of an appropriate test statistic only depends on the experimental design and on the type of the involved variables. We consider that the test statistic is correctly chosen with respect to the statistical context. The second step is the main concern of the following as the threshold on p-values can not be determined as in the univariate issue [12]. More particularly, the choice of the threshold influences the number of errors in tests decisions. For a given t, the number of possible errors in a multiple testing procedure are summarized in Table 1, with the same notations as in [2].

TABLE 1. *Numbers of errors in a multiple testing procedure.*

	declared non significant	declared significant	Total
H_0	U_t	V_t	m_0
H_1	T_t	S_t	m_1
Total	$m - R_t$	R_t	m

m is the known number of tested hypotheses. m_0 and m_1, respectively the number of true null and true alternative hypotheses, are unknown parameters. For a given threshold t for the p-values, R_t, the total number of significant tests, is an observed random variable. On the contrary, U_t and S_t on the one hand, and T_t and V_t on the other hand, respectively the number of right and wrong decisions, are unobserved random variables.
The univariate approach of test theory focus on optimal procedures, optimality being achieved when, while controlling the type-I error, the type-II error (when the test fails to reject a false null hypothesis) is minimized. Ideally, a multiple testing procedure would minimize both the number V_t of false positives (type-I errors) and the number T_t of false negatives (type-II errors). More false positives can occur when the number of tests increases: multiplicity necessitate to clearly define global type-I error rates, at the level of the whole set of tests instead of the level of individual tests. Naturally, extending the single test approach to multiple tests consists initially in controlling the risk of wrongly rejecting H_0 at least once, (called the Family-Wise Error Rate, FWER) considering $\mathbb{P}(V_t > 0)$. Controlling type-I error is of most importance in these contexts

 ISSN: 2102-6238

where a moderate number of tests are simultaneously performed, and the power issue is often set apart. Note that a definition of optimality does not arise and finding the best multiple testing procedure is still an open question [34].

Large scale multiple testing For the last two decades, innovative improvements have been made to face new scientific challenges. Particularly, high throughput technologies result in huge volume of data that allows the global analysis of complex systems. In these situations, the number of measured variables is close to several thousands, whereas the sample size is about some tens at most. Data are then said in high-dimension. At the root of the main issue presented in this article lies questions raised by the analysis of DNA microarray data. DNA microarrays are a biotechnology that allows the simultaneous measurement of gene expressions, at the level of the whole genome. Such data can be used, for example, to diagnose tumors, to profile drug-effect, or to group genes with similar expression patterns associated to common biological processes. This biological context has markedly contributed to the development of the statistical methodology for multiple testing in high dimensional data [16, 37, 13].
Indeed, the first step in the analysis of such data is called differential analysis. It aims at identifying the subset of differentially expressed genes *i.e* the subset of genes whose expression levels differ with respect to a covariate of interest, that can be either categorical, such as treatment/control status, or continuous such as a drug dose.
From a statistical point of view, the biological question of differential analysis is restated as a multiple hypotheses testing issue, considering the simultaneous tests for each gene of the null hypothesis H_0*: "there is no association between the expression levels and the covariate"*.
The context evoked previously induces thousands of simultaneous tests, one for each gene of the genome. Procedures, by Bonferroni [8] for example, controlling the FWER become highly conservative as the number of tests increases. As a matter of fact, the number of truly rejected null hypotheses (true positives) is very low.
Therefore, controlling the FWER has appeared unsuitable in a high-dimensional setting. An approach that has turned out to be much more appropriate in high dimension is to control the False Discovery Rate (FDR) [2], which is the expected proportion of false positives among the rejected hypotheses:

$$FDR_t = \mathbb{E}\left(\frac{V_t}{R_t}\right) \tag{1}$$

V_t/R_t being set to 0 when $R_t = 0$. This approach is useful in exploratory analyses, where one aims at maximizing the discoveries of true positives, rather than guarding against one or more false positives. Many methods have been proposed to control the FDR, the most famous being due to Benjamini and Hochberg [2] and called hereafter the BH procedure. The cut-off on the p-values under which the hypotheses are rejected is derived from the increasingly ordered p-values $p_{(k)}$ as follows: $t_\alpha = p_{(k^*)}$ with $k^* = \arg\max_k \left\{m\pi_0 p_{(k)}/k \leq \alpha\right\}$, provided the proportion of true null hypotheses $\pi_0 = \frac{m_0}{m}$ is known. In addition to being an interesting quantity in itself, for its interpretation, π_0 is a key parameter in assessing or controlling error rates. It has been recently shown [5, 21] that a more accurate estimation of π_0 would improve the power of multiple testing procedures. Generally, plugging-in an estimate of π_0 into the definition of the threshold on p-values corrects for the control level of the FDR and results in a less conservative procedure.

 ISSN: 2102-6238

Adaptative procedures, including π_0 estimation, has then emerged in the literature [40, 3]. See [24] for a comparative review of the main estimation methods of π_0.
Most of multiple testing procedures assume that the p-values are independently distributed according to a two-component mixture model [16], characterizing p-values distributions under the null hypothesis ($g_0 := \mathscr{U}_{[0;1]}$) and under the alternative one (g_1, unknown), respectively.

$$g(p) = \pi_0 g_0(p) + (1-\pi_0) g_1(p) \tag{2}$$

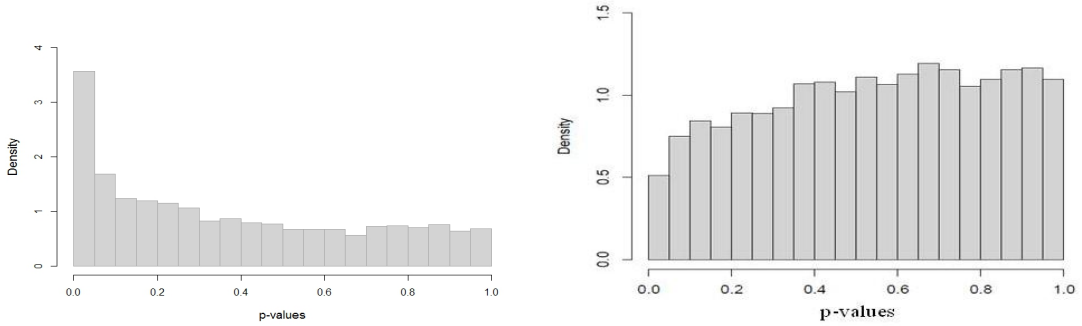
where $g_0(p) = 1, \forall p$. Further conditions are necessary to ensure identifiability of π_0, the mixing parameter of this model, which can be obtained for instance by assuming that g_1 is a decreasing function of the p-values with $g_1(1) = 0$ [19].
This setting is now illustrated with two examples of microarray data.

Example 1 (Breast Cancer study [20])**.** *These data were primarily analyzed in order to compare expressions of three types of breast cancer tumor tissues:* `BRCA1`*,* `BRCA2` *and* `Sporadic`*. The raw expression data, downloaded from* `http://research.nhgri.nih.gov/microarray/NEJM_Supplement/`*, initially consist of* 3 226 *genes in* 22 *arrays;* 7 *arrays from the* `BRCA1` *group,* 8 *from the* `BRCA2` *group and* 6 *from the* `Sporadic` *group. The label of one sample being unclear, it has been removed from the study. The same preprocessing procedure as in [39] is used:* 56 *genes presenting some suspiciously large expressions (larger than* 20*) are removed and the data are finally* $\log_2$ *transformed. The analysis focuses on the identification of differentially expressed genes among the* $m = 3\ 170$ *included in the study between the two types* `BRCA1` *(*7 *cases) and* `BRCA2` *(*8 *cases) using classical t-tests. Figure 1(a) displays the distribution of the raw p-values of these t-tests.*
Applying the BH procedure to determine which test should be rejected controlling the FDR, at a significance level $\alpha = 0.05$*, a list of* 96 *genes are declared differentially expressed.*

Example 2 (Lipid Metabolism study [26])**.** *These data, provided by the INRA Animal Genetics department in Rennes (France), describe chicken hepatic transcriptome profiles for* $m = 9\ 893$ *genes of* $n = 43$ *half-sib male chickens, selected for their variability on abdominal fatness (Af). The aim is to study the relationships between hepatic gene expressions and abdominal fatness [7] and to map quantitative trait loci (QTL) for abdominal fatness in chickens. Animals, marker genotyping, transcriptome data acquisition and normalization are described in [26]. The normalized microarray dataset is available in the* R *package* `FAMT` *[10], which implements the method presented later in this article. Figure 1(b) displays the distribution of the raw p-values from the t-tests for the significance of the correlation coefficient between genes expressions and the abdominal fatness. Applying the BH procedure results in no positive genes for a FDR control at level* $\alpha = 0.05$*.*

Microarray experiments involving genome-wide scans usually presuppose most of the genes to be null so that $\mathscr{U}[0;1]$ should fit the right side of the p-values histogram. In case of Example 1, this hypothesis seems to be borne out by the distribution observed on Figure 1(a). On the contrary, the shape of the histogram in the case of Example 2 on Figure 1(b) clearly shows an abnormal under-representation of the p-values in the neighborhood of 0 (and consequently an over-representation of the p-values close to 1). Indeed, if all the gene expressions were all truly under the null hypothesis, the p-values should be uniformly distributed on $[0,1]$ and the proportion

 ISSN: 2102-6238



(a) P-values' distribution for the t-tests comparing mean genes expressions in the two tumor types of the Breast Cancer data.

(b) P-values distribution for the t-tests for the significance of the correlation coefficient between genes expressions and the abdominal fatness of the Lipid Metabolism data.

FIGURE 1. *P-values distribution for Examples 1 and 2*

of observed p-values under 0.05 should be close to 0.05, provided the gene expressions are independent. This marked departure of the empirical distribution of p-values from the theoretical uniform distribution has been recently considered by some authors as the impact of a high amount of dependence among tests (see [15, 25, 18]). These effects have a significant impact on simultaneous hypotheses testing, and must be accounted for in test decisions.

1.2. Multiple testing and dependence

Among the topics in the literature on multiple testing in high-dimensional data, the assumption of independence on which most of these procedures are based is recently discussed. The study of the impact of dependence between the variables is of great interest as taking into account dependence casts doubt on multiple testing procedures as a whole. Two topics are mainly identified:

(a) *Controlling (type-I) error-rates when p-values are no longer independently distributed.*

Some papers have especially focused on the control of the FDR under various patterns of dependence. An important contribution to this point is the proof that the BH procedure still controls the FDR under assumption of a certain class of dependence called positive [4]. Other proposals extending the initial condition of the BH procedure have also been proposed later [40, 6]. In fact, the general message seems to be that, for a high amount of dependence, the BH thresholding method tends to over-control the FDR, leading to more conservative rules than expected under the assumption of independence. Consequently, this also means that dependence affects the power of the BH procedure and its stability. Some authors [39, 41] proposed to modify the test statistic and recent proposals suggest to modify the theoretical null distribution [14]. Modification of the p-values threshold as in [4] or adaptative BH procedures as in [3] have been proposed, leading again to more conservative rules than expected under the assumption of independence. The common point of all these approaches to deal with dependence consists in taking effect on one of the two steps of multiple testing procedures, namely (1) in the formation of the tests statistics or of their null-distribution for the calculation of p-values, or (2) when defining the p-values threshold. In each case, the focus is on the control of the type-I error rate.

 ISSN: 2102-6238

(b) *Taking into account dependence between test statistics by borrowing information across the variables rather than treating them as independent.*

More recently, a common idea in a few papers [15, 25, 18] is that dependence between test statistics should be taken into account through a latent dependence structure rather than treating variables as independent. Indeed, dependence between tests is directly deduced from dependence between the involved response variables. The true signal and several confusing factors are often observed at the same time and these factors lead to misleading conclusion on tests decisions. In microarray data analysis, dependence between gene expressions may comes from some biological gene interactions, in which the studied biological process is not necessarily involved but which impact the level of gene expressions as well. Technological bias can also affect gene expressions, even if some pre-processing treatments of the data such as normalization aim at limiting their impacts. All these uncontrolled and unobserved factors are referred hereafter as data heterogeneity components.

The following methodology is in the continuation of the former idea, an overall framework to deal with dependence, and this article focuses on the properties of multiple testing procedures under dependence. Section 2 presents the proposed framework based on a factor model for the dependence structure, introducing the Factor Analysis for Multiple Testing (FAMT) method and showing the improvement brought by this approach. The effects of dependence on multiple testing properties are investigated through a simulation study and providing some analytical results. The implementation of FAMT necessitates to determine the number of factors included in the model, and the estimation of its parameters. Section 4 describes our proposal, which matches the specific context of high-dimensional data. Finally, an application to the analysis of gene expressions data is presented in Section 5. The FAMT method is implemented in the R package `FAMT` [10], available on the R-project web site and on its own web site `http://famt.free.fr`.

2. A general framework to account for dependence

For $k = 1,\dots,m$, let Y_k denotes the kth response variable among m. In a high-dimensional frameworks, m can be much larger than the number n of independent observations of $Y = [Y_1, Y_2, \dots, Y_m]$. For each response Y_k, the link with p explanatory variables is explicitly defined by the following regression model:

$$Y_k = r_k(x) + e_k \qquad \forall k \in [1;m] := \mathscr{M} \tag{3}$$

where x is the $p-$vector of covariates, r_k is an unspecified regression function and e_k is a random error term. For $k \in \mathscr{M}_0 \subset \mathscr{M}$ with $\#\mathscr{M}_0 = m_0$, $r_k(x) = r_k^{(0)}(x)$, where $r_k^{(0)}$ is an arbitrary function of interest and for $k \notin \mathscr{M}_0$, $r_k(x) \neq r_k^{(0)}(x)$. Multiple testing aims at finding out the response variables for which the null hypothesis $\mathrm{H}_0^k : \ r_k(x) = r_k^{(0)}(x)$ is not true.

The present article focuses on t-tests because they are of major interest in various applied situations but the general conclusions are valid for other types of tests such as Fisher's analysis of variance tests for example. The test statistics are therefore defined as normalized estimations of linear contrasts $c'\theta_k$ and are denoted $T_k = \sqrt{n}c'\hat{\theta}_k/(\sigma_k\sqrt{c'S_x^{-1}c})$, where S_x denotes the empirical variance-covariance matrix of the explanatory variables. Under the null hypothesis, their distribution is known and the p-value for each test is denoted p_k.

 ISSN: 2102-6238

Proposition 1.

$$\mathbb{E}(T_k) = \tau_k = \frac{\sqrt{n}c'\theta_k}{\sigma_k\sqrt{c'S_x^{-1}c}} \quad \mathbb{V}(T_k) = 1 \quad Cov(T_k, T_{k'}) = \rho_{kk'}, k \neq k'$$

Note that τ_k *equals* 0 *if* $k \in \mathscr{M}_0$

Proposition 1 shows that the correlation structure between the test statistics is directly inherited from the correlation ρ between the response variables. Note that this property is generally not true for other types of tests. Therefore, specific relationships between both correlation structures should be taken into account in order to adapt the following results to other testing procedures. More generally, dependence among the p-values is also straightforward inherited from dependence among the data.

In regression models such as (3), the residuals are usually assumed to be independent. In practice, and especially in gene expression data for example, unmodeled and/or uncontrolled factors can interfere with the true signal and then generate dependence across the variables. The consequence is that the residuals of model (3) are not independent, which violates the assumption of p-values distribution as in (2). In many areas, dependence can be explained by an underlying structure of unobserved factors, previously refered to as data heterogeneity. Our proposal consists in capturing data heterogeneity by modeling this latent structure. Residuals in models (3) are then split into two terms, one associated to the heterogeneity through latent variables Z, and independent residuals:

$$Y_k = r_k(x) + Zb_k' + \varepsilon_k, \tag{4}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ is a random vector with independent components. The mixed-effects regression models (4) are equivalently defined as fixed-effects regression models as in (3) but where residual variance Σ can be decomposed into the sum of two components: a diagonal matrix Ψ of specific variances $\psi_k^2 = \mathbb{V}(\varepsilon_k)$ and a common variance component $B'B$, where the kth row of B is b_k:

$$\Sigma = BB' + \Psi \tag{5}$$

This general approach is proved in the following proposition [25] which defines a general framework for multiple testing dependence.

Proposition 2 (see [25])**.** *Under assumption* (3)*, suppose that for each* ε_k*, there is no Borel measurable function* g *such that* $\varepsilon_k = g(\varepsilon_1, \ldots, \varepsilon_{k-1}, \varepsilon_{k+1}, \ldots, \varepsilon_m)$ *almost surely. Then, there exists a random* $Q-$*vector* Z*, with* $0 \leq Q \leq m$ *and, for all* $k = 1, \ldots, m$*, there exist* $Q-$*vectors* b_k *such that,*

$$Y_k = r_k(x) + Zb_k' + \varepsilon_k$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ *is a random vector with independent components.*

Model (4) establishes the existence of Q latent variables Z which capture the dependence among the variables in a $Q-$dimensional linear space. Without loss of generality, in the following, it is assumed that the latent variables Z have means 0 and variance I_Q. Furthermore, the estimation of the Factor Analysis model parameters as detailed in Section 4.1 assumes that the common factors

 ISSN: 2102-6238

are normally distributed. Therefore, model (4) can be viewed as a Factor Analysis model [27] and the latent variables are hereafter called (common) factors. Factor Analysis is an analytic tool used for many years in economics, social sciences and psychometrics, originally in the field of intelligence research [35]. It has only appeared recently in the study of the dependence structure in high dimensional datasets provided by microarray technology [31, 23]. The common point between all these applied sciences is to deal with large quantities of data. In such a case, we often need to determine a smaller set of synthetic variables that could explain the original set.
The following proposition gives the conditional distribution of the tests statistics $T = [T_1; \ldots; T_m]$, given the factors.

Proposition 3. *Conditionally on Z, T is normally distributed with, for* $k = 1, \ldots, m$, $\mathbb{E}(T^{(k)} \mid Z) = \tau_k + b_k' \tau_Z / \sigma_k$, *where* τ_Z *is the* $Q-$*vector defined by* $\tau_Z = \sqrt{n} c \hat{\theta}_z / \sqrt{c' S_x^{-1} c}$ *and* $\hat{\theta}_z$ *denotes the least-squares estimator of the* $p \times Q$ *matrix of the slope coefficients in the multivariate regression of Z onto the explanatory variables x. Moreover,* $\mathbb{V}(T|Z) = diag(\psi_k^2/\sigma_k^2)$. *Note that* τ_Z *is normally distributed with mean* 0 *and variance* $\mathbb{I}_Q$

Factor-adjusted test statistics $\tilde{T}_k$ are now defined as conditionally centered and scaled versions of the classical test statistics T_k:

$$\tilde{T}_k \quad = \quad \frac{\sigma_k}{\psi_k}\left[T_k - \frac{b_k'}{\sigma_k}\tau_Z\right]$$

The following proposition gives the distribution of $\tilde{T} = [\tilde{T}_1; \ldots; \tilde{T}_m]$.

Proposition 4. *Under assumption of a decomposition of the covariance matrix as in* (5), $\tilde{T}$ *is normally distributed with, for all* $k \in [1;m]$, $\mathbb{E}(\tilde{T}_k) = \tau_k / \sqrt{1 - h_k^2}$, *where* $h_k^2 = b_k b_k' / \sigma_k^2$ *is the communality of* Y_k. *Moreover,* $\mathbb{V}(\tilde{T}) = \mathbb{I}_m$.

The non-centrality parameter of $\tilde{T}_k$ being always larger than τ_k, the variable-by-variable power of the factor-adjusted tests are larger than for the t-tests. Furthermore, this non-centrality parameter, and consequently the power of the factor-adjusted tests, are increasing functions of the communality h_k^2, which confirms the idea that the multiple testing procedure can be improved by a correction of the individual test-statistics regarding their contribution to the common variability across variables. On the contrary, if the k^{th} variable does not contribute to the factor structure, $b_k = 0$ and $\tilde{T}_k$ coincides with the usual test statistic T_k. As the factor-adjusted tests statistics are independent, the associated factor-adjusted p-values are also independent [17].
Estimated factor-adjusted test statistics $\tilde{T}$ are obtained by plugging estimates of the factor model's parameters in the test statistic. In Section 4, ML estimates are proposed. As these estimators of the variance parameters are consistent, this does not affect the asymptotic distribution of the factor-adjusted test statistics. By analogy with the classical situation, we propose to take into account the effect of estimating the variance parameters in small-sample conditions by approximating the null distribution of $\hat{\tilde{T}}_k$ by a Student distribution.
Note that the factor-adjusted test statistics can be equivalently obtained by computing the usual test statistics on the data centered with respect to the dependence kernel ZB':

$$\tilde{Y}_k = Y_k - Zb_k' = r_k(x) + \varepsilon_k \tag{6}$$

 ISSN: 2102-6238

3. Impact of dependence...

The impact of dependence on multiple testing procedures and improvements brought by the general framework introduced in previous section are illustrated by simulation studies and some analytical results are also provided in this section. Model estimations issues are set apart in this section and are developed in Section 4. In the following simulation studies, the parameters estimations are computed using the method presented in Section 4.
In similar simulation studies presented in the literature, the patterns of correlation for simulated data is relatively simple: equi-correlated data [1] or block equi-correlated data [22] for example. These dependence structures are simple to implement and easily interpretable. Controlling the level of dependence within each scenario is possible by varying the value(s) of the correlation matrix. Nevertheless, this modeling of the dependence is far from reality, mainly in the analysis of microarray data, where the connections between variables of interest are sometimes much more complex. In the following studies by simulations, in a manner similar to [21], we propose to consider a set of correlation matrices, while imposing a constraint of conditional independence. The theoretical variance-covariance matrix for the simulated datasets is split into two components, a diagonal matrix Ψ of specific variances and a common-variance component $B'B$: $\Sigma = BB' + \Psi$. Conditionally on the common structure, the data are independent. The desired level of dependence is reached by weighting the common structure (BB') by a coefficient from 0 (independence) to 1 (highest level of dependence).

3.1. ... on p-values distribution

Model (2) assumed both independence of the p-values and uniformity of the null component g_0. In some situations of weak correlations, or "clumpy" correlations (many groups made of a small number of variables with high correlation within groups and no correlation between groups [38]), the uniform assumption still holds [24]. However, in the presence of highly dependent data, g_0 can severely deviate from uniform distribution. Considering the factor-adjusted data to compute the p-values redresses the p-values distribution with respect to the raw data case, under dependence. This point is illustrated by simulation scenarios with increasing amounts of dependence among data.

Simulation study 1. *10 levels of dependence are considered, from independence (scenario* 0*) to highly correlated data (scenario* 9*). The proportion* $tr(BB')/tr(\Sigma)$ *of common variance increases along with the scenarios:*

TABLE 2. *Common variability (%) for the 10 simulated scenarios*

Scenario	0	1	2	3	4	5	6	7	8	9
Common variability (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19

In each scenario, 1 000 datasets are simulated according to a multivariate normal distribution: each dataset is composed of $m = 500$ *variables and* $n = 60$ *observations such as* $Y_{n\times m} \sim \mathcal{N}_m(0;\Sigma)$. *Besides, let's consider a binary variable X such that the observations are split into two arbitrary groups of size* $n/2$. *The multiple testing procedure aims at finding out which variables have different expectations in two groups with equal sample size. For each dataset, the p-values of the*

 ISSN: 2102-6238

usual t-tests for the comparison of means are calculated, both on raw and factor adjusted data. In this simulation study, for each dataset, all hypotheses are true null, so that all set of p-values should be uniformly distributed.

Distribution of null p-values under dependence Figure 2 reproduces the average histograms of p-values with 95% error bars in situations of independence, intermediate level of dependence and high level of dependence for raw data (Figure 2(a)) and factor-adjusted data (Figure 2(b)). Obviously, this shows that the assumption of uniformity for the null p-values distribution is true on average (grey histograms), whatever the level of dependence, for both raw and factor-adjusted data. However, in case of dependent data, the distribution of raw p-values can show marked departures from uniformity. This leads either to a much larger representation of the p-values close to 0 (and consequently an under-representation of the p-values close to 1, the distribution of null p-values being decreasing instead of being flat) or inversely much lesser small p-values than expected under uniformity (the distribution of null p-values being increasing instead of being flat). It should be noticed that the first situation is much more marked. This point has a direct consequence on the proportion of false positives in tests decisions, which can be much higher than expected under the uniform assumption for p-values distribution.
For each scenario, Table 3 gives the proportion of significant Kolmogorov-Smirnov goodness-of-fit tests, the null hypothesis of each test being that the null p-values distribution is uniform (level of significance of 5%). As the dependence structure gets stronger, going from scenario 0 to scenario 9, the proportion of significant tests increases, up to 80% for the highest level of dependence. This violation of the uniformity of the null distribution is also mentioned in [15], which reports that correlation can widen or narrow down the distribution of test statistics with respect to the theoretical null distribution.

TABLE 3. *Proportion of significant goodness-of-fit tests for uniformity of null p-values within each scenario of simulation (Kolmogorov-Smirnov tests; significance level:* 5%*) - case of raw-data*

Scenario	0	1	2	3	4	5	6	7	8	9
Prop. of sig. tests (%)	4.4	7.1	20.2	38.0	56.1	63.8	68.5	71.6	75.5	80.1

Considering the factor-adjusted p-values, there is low variability around this uniform distribution, whatever the level of dependence, as suggested by the 95% error-bars on Figure 2(b).

3.2. ... on error-rates

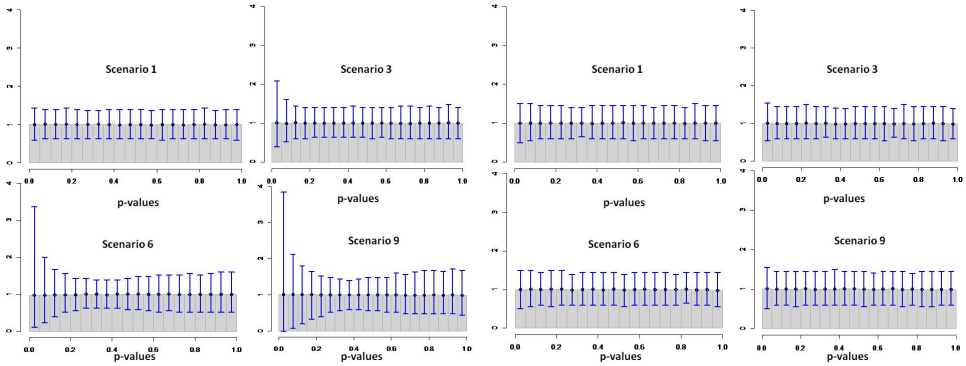
Variance of the number of False Positives (V_t) For a given threshold t, V_t is defined as the number of erroneous rejections of the null hypotheses. For independent test statistics, V_t is distributed according to a binomial distribution: $V_t \sim \mathscr{B}in(m_0,t)$. This random variable has mean $m_0 t$ and variance $m_0 t(1-t)$. Under general dependence, the following proposition holds:

Proposition 5 (Variance of the number of false-positives (V_t))**.**

$$\mathbb{E}(V_t) = m_0 t \tag{7}$$

$$\mathbb{V}(V_t) = \left[m_0 + \sum_{k \neq k' \in \mathscr{M}_0} D_t(\rho_{kk'})\right] t(1-t) \tag{8}$$

 ISSN: 2102-6238



(a) P-values computed from raw data (b) P-values computed from factor-adjusted data

FIGURE 2. *Mean histograms of the p-values for 1 000 simulations with 95% error bars, for 4 dependence structures (low (1), intermediate (3 and 6) and high (9) level of dependence)*

$D_t(\rho_{kk'})$ is defined as follows:

$$D_t(\rho_{kk'}) = \frac{\sum_{k=1}^{m}\sum_{k'=1,k'\neq k}^{m} G_{kk'}(t) - G_k(t)G_{k'}(t)}{m(m-1)}. \tag{9}$$

Where, G denotes the distribution function for the kth p-value, such as for the k^{th} variable: $G_k(t) = \mathbb{P}(p_k \leq t)$ and for $k \neq k'$, the bivariate distribution function is $G_{kk'}(t) = \mathbb{P}(p_k \leq t; p_{k'} \leq t)$. Figure 3 shows that, for any preset t, $D_t(\rho)$ is a U-shaped function, closed to an equivalent term appearing in Owen's formula [29] for the variance of the number of false discoveries and in [15], where the bivariate normal probability function is also involved in the expressions of the variance inflation. $D_t(\rho)$ equals zero for independent variables, and increases as the correlation grows.

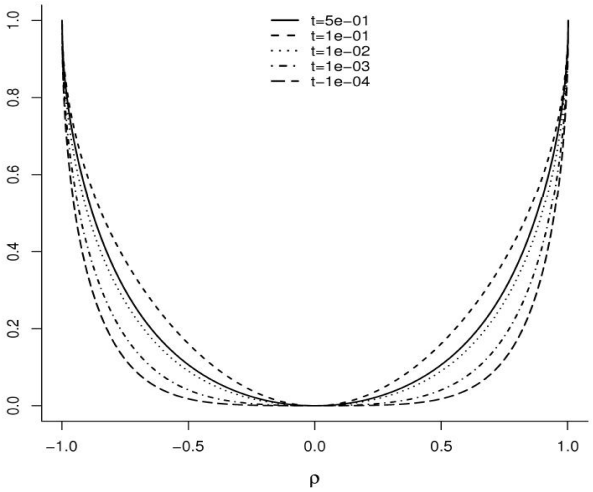


FIGURE 3. $D_t(\rho)$ *for various values of the threshold t along with the correlation* ρ

 ISSN: 2102-6238

Therefore, by comparison with the binomial variance $m_0 t(1-t)$ for the independent case, the impact of dependence on the variance of V_t can be measured by the term $\sum_{k \neq k' \in \mathscr{M}_0} D_t(\rho_{kk'})$, where $\rho_{kk'}$ is the correlation between the test statistics for variables k and k'. Proposition 5 shows that correlation modifies the distribution of V_t by increasing its dispersion, leaving its expectation unchanged. Correlation has an important impact on the tail of the distribution, which is directly involved in the calculation of the error rates. Proposition 5 confirms that a strong correlation structure leads to a very unstable distribution of the number of false discoveries.

False Discovery Rate and Non Discovery Rate Proposition 5 shows that correlation modifies the distribution of V_t by increasing its dispersion, leaving its expectation unchanged. Therefore, correlation shall have an important impact on the tail of the distribution, which is directly involved in the calculation of the error rates.

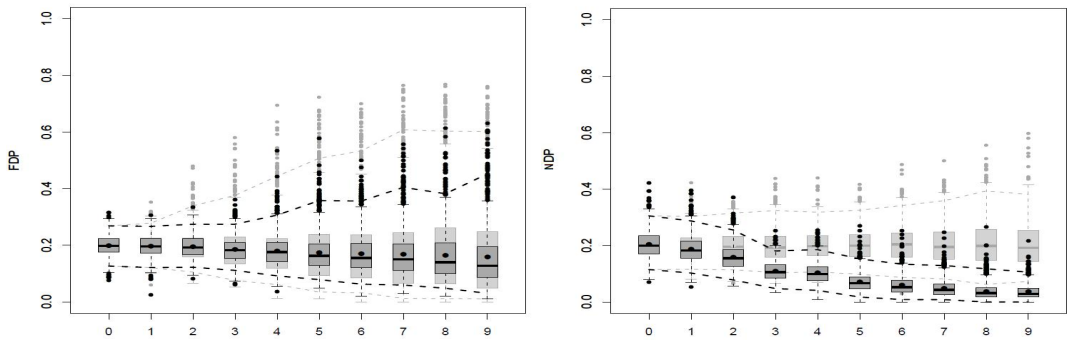
Simulation study 2. *The same simulation scheme as in the first simulation study is considered here. In addition, for* $m_1 = 100$ *variables, expectations in each group A and B are set so that the usual t-tests have a variable-by-variable power of* 0.8. *For the remaining* $m_0 = 400$ *variables, the difference is set to* 0. *For each dependence level (see Table 2), 1 000 datasets are simulated according to a normal distribution. Finally, for each dataset, the p-values of the usual t-tests for the comparison of means are calculated, both on raw and factor adjusted data. The BH procedure is used to define the threshold* t *on p-values, with* $\alpha = 0,2$. *The threshold on the p-values is determined using the* R *package* `multtest` *[30]. The true False-Positives Proportion (FDP) and the true Non-Discovery Proportion (NDP) are then computed for each dataset.*

Figure 4 reproduces multiple boxplots of the distributions of the FDP and NDP, for both raw data (gray) and factor-adjusted data (black). Let's first consider the case of raw data: mean of FDP, which is FDR, is steady for all scenarios, but its variability sharply increases along with the proportion of common variability (Figure 4(a)). Figure 4(b) shows that the fraction of common variance generates slight instability in the distribution of NDP_t. The mean NDP_t, which can be viewed as a type-II error rate, remains steady whatever the level of dependence. By comparison, the distribution of the FDP is clearly stabilized when considering the factor-adjusted strategy: using the usual t-tests, the variability of the FDP reaches up to four times the variability obtained under independence whereas it remains controlled at almost the same level using the factor-adjusted method. Another striking property of our method is the very important improvement of the global power of the multiple testing procedure compared with the BH procedure based on t-tests, illustrated by Figure 4(b). This result probably illustrates the idea that dependence between the responses should not just be seen as a nuisance for controlling the FDR but also as a support to provide improved estimation of the effects of covariates.

False Discovery Rate and False Discovery Proportion Multiple testing theory usually focus on the control of type-I error rates, such as FDR control. The present section illustrates issues occurring in the estimation of the FDR in case of dependent data.
The empirical FDR estimator is defined as

$$\widehat{FDR_t} = \frac{m_0 t}{R_t} = \frac{\mathbb{E}(V_t)}{R_t} \tag{10}$$

 ISSN: 2102-6238



(a) FDP (b) NDP

FIGURE 4. *False Discovery Proportion (FDP) and Non Discovery Proportion (NDP), along with the 10 scenarios in Simulation study 2: raw data (grey) and factor adjusted data (black)*

Let's introduce a novel estimate for the FDR, which takes advantage of the factor structure, and compare it with the empirical estimate defined in (10) thanks to the data of Simulation study 2. The following expression gives the expectation of V_t conditionally on the factors.

$$\mathbb{E}(V_t|Z) = \sum_{k\in\mathscr{M}_0} \mathbb{P}(p_k \leq t|Z) = \sum_{k\in\mathscr{M}_0} G^Z(k,t) \tag{11}$$

The above expression of $\mathbb{E}(V_t|Z)$ is now used to define a conditional estimate $\widehat{FDR}_t^Z$ of the FDR, by analogy with the proposition made by [15], who defines FDR_t^A as $\mathbb{E}(V_t|A)/R_t$, where A is a random variable which value essentially differs according to the amount of dispersion among the correlations between the test statistics.

$$\begin{aligned} \widehat{FDR}_t^Z &= \frac{\mathbb{E}(V_t|Z)}{R_t} = \frac{m_0 t}{R_t} + \frac{\sum_{k\in\mathscr{M}_0} G^Z(k,t)}{R_t} - \frac{m_0 t}{R_t} \\ &= \widehat{FDR}_t . \left[1 + \frac{\sum_{k\in\mathscr{M}_0}(G^Z(k,t)-t)}{m_0 t}\right] \end{aligned} \tag{12}$$

This FDR estimate is defined as a correction of the unconditional estimate, accounting for the correlation among the test statistics through factors.

The conditional and the empirical estimates of the FDR are now compared on the simulated datasets of the Simulation study 2. For each dataset and for $t = 0.05$, the conditional estimate FDR_t^Z is estimated, together with the empirical estimate $\widehat{FDR}_t$. To avoid discussions about the impact of the estimation of m_0 in this comparative study, $m_0 = 400$ is supposed to be known. For each scenario, Table 4 gives the regression coefficients between the observed false discovery proportion FDP_t and both FDR estimations. Results for scenario 9 are illustrated on Figure 5 by plots of both FDR estimates versus FDP_t.
Ideally, a correlation of 1, or at least a high and positive correlation, between the estimated FDR and the true FDP is expected. In practice, this means that the estimation of the FDR is a suitable indicator to accurately reflects the true proportion of false positives.

 ISSN: 2102-6238

TABLE 4. *Regression coefficients between FDR estimates and FDP*

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_t$	-0,197	-0,196	-0,189	-0,176	-0,187	-0,171	-0,175	-0,182	-0,169	-0,171
FDR_t^Z	-0,197	0,025	0,578	0,815	0,844	0,886	0,907	0,889	0,915	0,913

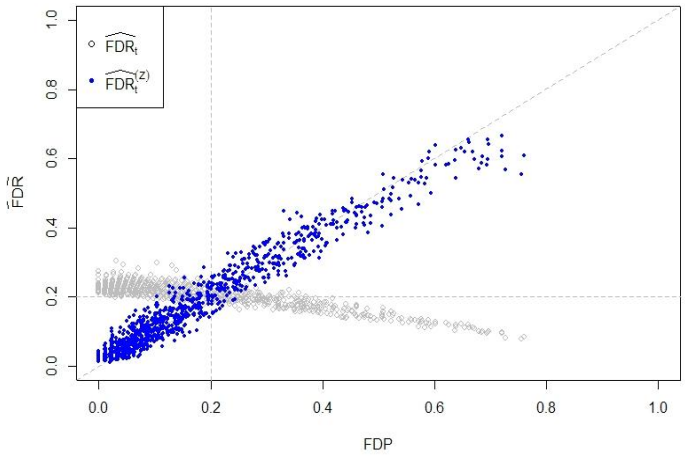


FIGURE 5. *Estimated FDR versus observed true proportion of false positives (FDP) with* $\alpha = 0.2$ *for a high level of dependence (scenario 9 in Simulation study 2)*

The unconditional estimate $\widehat{FDR}_t$ is negatively correlated with the observed FDP_t, which can result in strongly misleading estimations especially when FDP_t is high. Figure 5 shows that this concern is particularly clear for large fractions of shared variance (scenario 9). For small fraction of shared variance, FDR_t^Z suffers from the same problem, essentially because the number of factors is most often estimated by zero, in which case $\widehat{FDR}_t^Z = \widehat{FDR}_t$. From scenario 2 to 9, when the factor structure is clearer, $\widehat{FDR}_t^Z$ is positively correlated with FDP_t, confirming the importance of accounting for the correlation between test statistics in multiple testing procedures.

4. Factor Analysis in high-dimension

The method presented in Section 2 involves the estimation of the model parameters. Several approaches could be planed. As an example, the Surrogate Variable Analysis [25] which is based on a similar model decomposition as in (4) considers a singular values decomposition. On our side, we consider that models (4) can be viewed as a Factor Analysis model. Therefore, this section focus on the practical aspects of implementing Factor Analysis, in a high-dimensional setting. More precisely, we now consider two issues in carrying Factor Analysis in high-dimension: factor number determination and parameter estimation. For each issue, different methods are presented and the choice the most appropriate one for the present context of multiple testing is discussed.

 ISSN: 2102-6238

4.1. Estimation of the model parameters

Let's consider that each variable Y_k, $k \in [1..m]$, can be expressed as a linear combination of common factors, Z_q, $q \in [1..Q]$ and a specific term as in (4). The issue is then to determine $\hat{B}$ and $\hat{\Psi}$, respectively the estimation of the loadings, representing the weight of the considered variable on the Z structure, and the estimation of the uniquenesses, so that: $S = \hat{B}\hat{B}' + \hat{\Psi}$.

Existing methods There are several methods to extract factors from the data. Generally, two of them are implemented in the leading statistical software: Principal Factoring (PF) and Maximum Likelihood (ML). The ML method has some nice statistical properties such as asymptotic efficiency, invariance under change of scale, and the existence of a test for additional factors. ML method is favored in the following for these interesting properties in an inferential aim [32].

ML Factor Analysis for high-dimensional data To avoid "Heywood cases" (specific variances greater than 1) that can be brought by Newton-Raphson algorithms in the seek of the ML estimators, an EM algorithm is proposed [33]. This class of algorithm is now a very popular tool for iterative ML estimation in issues involving missing or incomplete data. In the Factor Analysis framework, we aim at estimating the parameters of a multivariate normal model with missing data, where in this case, the missing data are the unobserved latent variables Z, which are assumed to be normally distributed. The EMFA algorithm is described hereafter:

1. **E-step**: Scores estimation

$$\begin{aligned} \mathbb{E}(Z^{(i)}|Y^{(i)}) &= Y^{(i)}\Psi_0^{-1}B_0(\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} \\ \mathbb{E}(Z^{(i)}Z'^{(i)}|Y^{(i)}) &= (\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \end{aligned}$$

2. **M-step**: Estimation of B and Ψ

$$\begin{aligned} B_1 &= \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1} \\ \Psi_1 &= diag\left[S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B_1'\right] \end{aligned}$$

A bias correction in order to account for the small sample conditions in which these models are usually estimated is proposed [18]. EM Factor modeling can indeed be viewed as a particular nonlinear smoothing procedure where $H_z = \hat{Z}\left[\sum_{i=1}^{n} S_i^{(z)}\right]^{-1}\hat{Z}'$ stands for the smoothing matrix of the factor model. c_z denotes the trace of $I_n - H_z$ and the degree-of-freedom corrected estimator $(n/c_z)\hat{\Psi}$ of Ψ is deduced.

4.2. Determination of the number of factors

In Factor Analysis, the first step consists in estimating the number of factors Q to be considered in the model. This step is the most crucial in conducting Factor Analysis and must balance between

 ISSN: 2102-6238

parsimony (low model complexity), and accuracy in representing the correlation structure.Under-estimation of Q leads to loss of information by ignoring a factor or recombining one with another: the loadings for measured variables are therefore biased and interpretation based on them would not be much reliable, the true structure of the data being concealed. Over-estimation is commonly considered as less severe, but considering too many factors underlines minors factors: interpretation is difficult and such model is unlikely to be robust for replication. Therefore, considering both too few and too much factors has significant consequences in the reduction of information, affecting parameters estimation and data interpretation. Because of all these reasons, the number of factors issue leads to plenty of methods proposed in the literature, with more or less subjective decision rules.

Existing methods Many popular methods can be implemented to achieve a good compromise between the specific and common variance components of the model, such as parallel analysis [28] or the scree test [9]. Nevertheless, there is no consensus in which method is more appropriate, mainly in a high dimensional context.

A criterion based on the variance inflation of the false positives In the present multiple testing framework, in which an overestimation of the common variance part can result in a higher false discovery proportion, we propose a new criterion to determine the number of factors to retain, which matches with high-dimension constraints. It consists in the minimization of an ad-hoc criterion, which can be viewed as the variance inflation of the number of false discoveries due to dependence. This variance inflation is deduced from expression (9), where the sums are restricted to indices of variables in $\mathscr{M}_0$ [18]. This criterion is successively estimated using the residual matrix obtained with an increasing number of factors. Let's consider the following model defined previously in (4) and assuming q common factors for each variable k: $Y_k = \mu_k + Zb_k'^{(q)} + \varepsilon_k^{(q)}$. If the whole correlation structure is well modeled by the Factor Analysis model, then the residual correlation should be zero. Indeed, $cov(\varepsilon_k;\varepsilon_{k'}) = \sigma_k\sigma_{k'}\rho_{kk'} - b_k b_{k'}'$ and $\mathbb{V}(\varepsilon_k) = \Psi_k$ so the residual correlation is defined, assuming a q-common factors model, by:

$$\rho_{kk'}^{(q)} = \frac{\sigma_k\sigma_{k'}\rho_{kk'} - b_k^{(q)}b_{k'}'^{(q)}}{\sqrt{\Psi_k\Psi_{k'}}} \tag{13}$$

Beforehand, let's recall the definition a U-shaped criterion called $D_t(\rho_{kk'})$ which appears in the expression of the variance of the number of false positives. It ranges from 0 in $\rho = 0$ to 1 in $\rho = -1$ and $\rho = 1$. Considering the $D_t(\rho_{kk'})$ criterion for each pair of variables $\{Y_k, Y_{k'}\}, k \neq k' \in \mathscr{M}_0$, the proposed method to determine the number of factors is to choose Q satisfying:

Proposition 6. *If $\rho_{kk'}^{(q)}$ is the residual correlation between Y_k and $Y_{k'}$ as in (13), let's define Q as:*

$$Q = \underset{q\in[0:q_{max}]}{\operatorname{argmin}} \frac{1}{m(m-1)} \sum_{k\neq k'\in[1;m]} D_t(\rho_{kk'}^{(q)}) \tag{14}$$

Q is the number of factors that minimizes the mean of $D_t(\rho)$ criterion over all the pairs of variables. In the multiple testing context, this criterion calculated over variables in $\mathscr{M}_0$ allows to extract the number of factors that minimizes the variance of V_t (8). The consequence is that multiple

 ISSN: 2102-6238

testing procedures are stabilized. In practice, this criterion is successively estimated using the residual matrix obtained with an increasing number of factors and the retained number of factors is obtained when the variance inflation is minimized.
This procedure for the number of factors determination is now implemented for the 10 scenarios introduced in Simulation study 1, in which the true number of factors is $Q = 5$. Figure 6 reproduces barplots of the distribution of $\hat{Q}$. It clearly shows that when the proportion of common variance is small, the estimated number of factors is relevantly lower than Q and when the factor structure dominates the specific part, $\hat{Q}$ provides a precise estimation of Q.

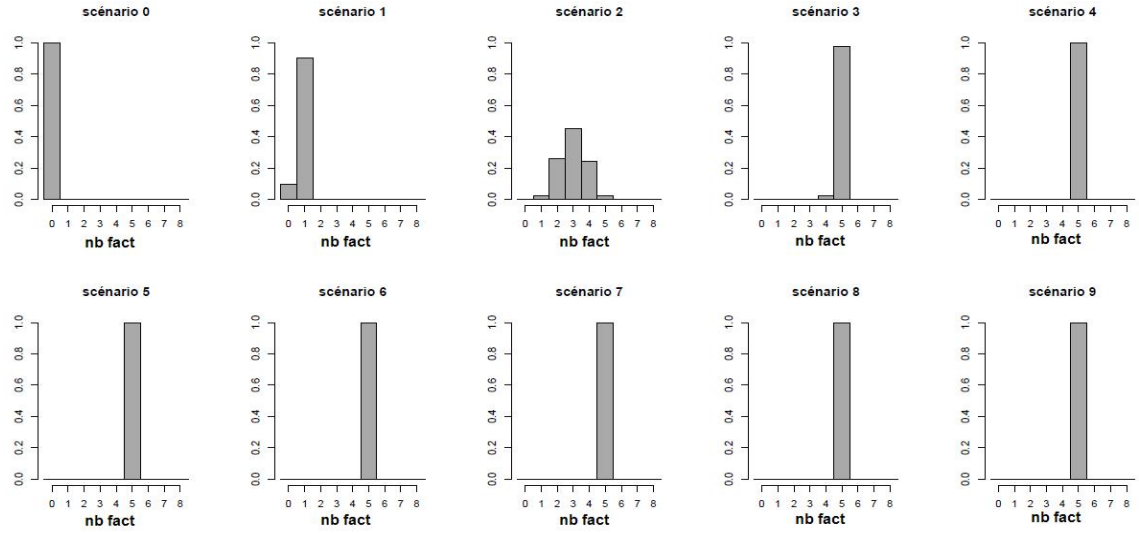


FIGURE 6. *Distributions of the estimated number of factors along with the dependence level. From scenarios 4 to 9, $\hat{Q}$ turns out to be constant and equal to Q=5.*

In practice, implementing this method requires to calculate the $D_t(\rho_{kk'})$ criterion for all pairs of variables, namely $m \times (m-1)$. As the correlation matrix is symmetric, it actually requires to consider $m \times (m-1)/2$ pairs. In the multiple testing framework, this is reduced to $m_0 \times (m_0-1)/2$ pairs, as the factors must be extracted from the variables in $\mathscr{M}_0$. Indeed, the information about the experimental condition X should not be captured by the common factors. The choice of subset of variables considered in $\mathscr{M}_0$ is discussed in section 6. In any case, the number of pairs is very huge, and can reach several thousands. As also suggested by [29], we propose to consider a range of η values, equi-distributed on $[0;1]$. Then, we count the number of times each value appears in the correlation matrix, considering an approximation of correlations by taking their absolute value and rounding them at the specified significant decimal figure:

$$\sum_{k \neq k'} D_t(\rho_{kk'}) \approx \sum_{j=1}^{\eta} n_j D_t(\rho_j) \tag{15}$$

where n_j represents the number of variables pairs for which the rounded correlation is equal to ρ_j. If the correlation distribution is symmetric, which is the case in most of applications, then the approximation given in (15) is accurate and leads to a sharp increase in computation time.

 ISSN: 2102-6238

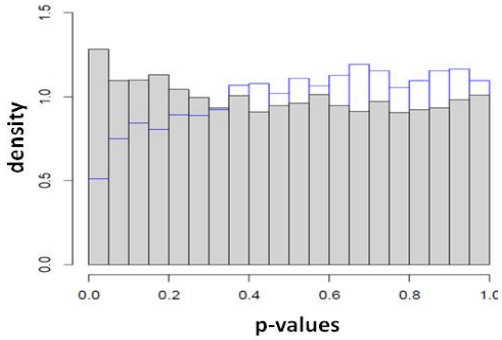
In the R package FAMT, the minimized criterion is based on the following approximation:

$$\frac{1}{m(m-1)}\sum_{k\neq k'} D_t(\rho_{kk'}) \approx \int_{[0;1]} D_t(\rho_{kk'}) f(\rho) d\rho$$

where $f(\rho)$ is the null correlations density function. It is estimated thanks to sampling in observed correlations.

5. Application to genomics

The methodology presented in Section 2 is now applied on data introduced in Example 2. Microarray experiments were initially conducted to identify differentially expressed genes from hepatic transciptome profiles for $m = 11\,213$ genes of $n = 45$ half-sib male chickens variable for abdominal fatness (AF) ([26], [7]). The data are provided by the INRA Animal Genetics department in Rennes, France.



(a) Histograms of raw p-values (empty bars) and factor-adjusted p-values (grey bars)

	Individual information			Gene information			
	hatch	dam	weight	oligo size	chip block	chip row	chip column
Factor 1	**8.92E-05**	0.139	0.129	**2.20E-16**	**2.20E-16**	0.074	0.179
Factor 2	0.074	0.913	**4.70E-03**	**2.20E-16**	**2.20E-16**	0.041	0.857
Factor 3	**1.90E-02**	0.848	0.489	**2.55E-14**	**2.20E-16**	0.716	0.376
Factor 4	**6.00E-03**	0.127	0.959	**1.41E-07**	**2.20E-16**	0.707	0.167
Factor 5	0.435	0.217	0.884	0.529	**2.20E-16**	**4.97E-03**	**9.99E-05**
Factor 6	0.946	0.412	0.615	**1.79E-07**	**2.20E-16**	0.876	**5.11E-07**

(b) Interpretation of the common factors with respect to heterogeneity components [7]: p-values from the test of the link between each factor and some individual or gene information

FIGURE 7. *FAMT on the Lipid Metabolism study of Example 2*

In order to figure out the differences between both analyses (from raw and factor-adjusted data), Figure 7(a) compares the empirical distributions of the raw and the factor-adjusted p-values. As noticed previously, the shape of the raw p-values distribution clearly shows an abnormal under-representation of the p-values in the neighborhood of 0. Indeed, if all the gene expressions were all truly under the null hypothesis, the p-values should be uniformly distributed on $[0, 1]$ and the

proportion of observed p-values under 0.05 should be close to 0.05, provided the gene expressions are independent. This marked departure of the empirical distribution of of p-values from the density function of a uniform distribution has been recently considered by some authors as the impact of a high amount of dependence among tests [15, 25, 18]. Applying the BH procedure on the raw p-values results in no positive genes. Factor-adjustment restores independence between tests statistics, which results in a correction of the distribution of the p-values from the concave shape observed on Figure 7(a). Indeed, it seems that a large amount of p-values are uniformly distributed and a few small p-values shall correspond to significant genes. The FAMT method is conducted considering a model with $Q = 6$ common factors, which corresponds to the minimum value of the variance inflation criterion introduced in Proposition 6. The model parameters are estimated with this choice of a 6-factor structure and π_0 is estimated using a smoothing spline method [36] applied on the factor-adjusted p-values. As the factors are designed to be independent from the explanatory variables (the abdominal fatness), they shall be described according to the other available covariates such as the hatch, the total body-weight of the chicken, gene (length) or microarray information (block, column, row) as shown in Table 7(b). This confirms *a posteriori* the strength of capturing heterogeneity through latent variables in the model rather than correcting multiple testing procedures at each step.

For a threshold of $t = 0.05$ on the p-values (no correction for multiplicity), $d_1 = 287$ and $d_2 = 688$ gene expressions are declared significantly correlated to the abdominal fatness trait, considering the raw and factor-adjusted data respectively. The list of significant tests is larger in the case of factor-adjusted data but also, and above all, it is much more relevant as illustrated by Figure 8 which represents the factorial maps obtained from PCA on each list respectively. On the right map, the discrimination between lean (L) and fat (F) chicken is much more highlighted along with the first axis than on the left map.

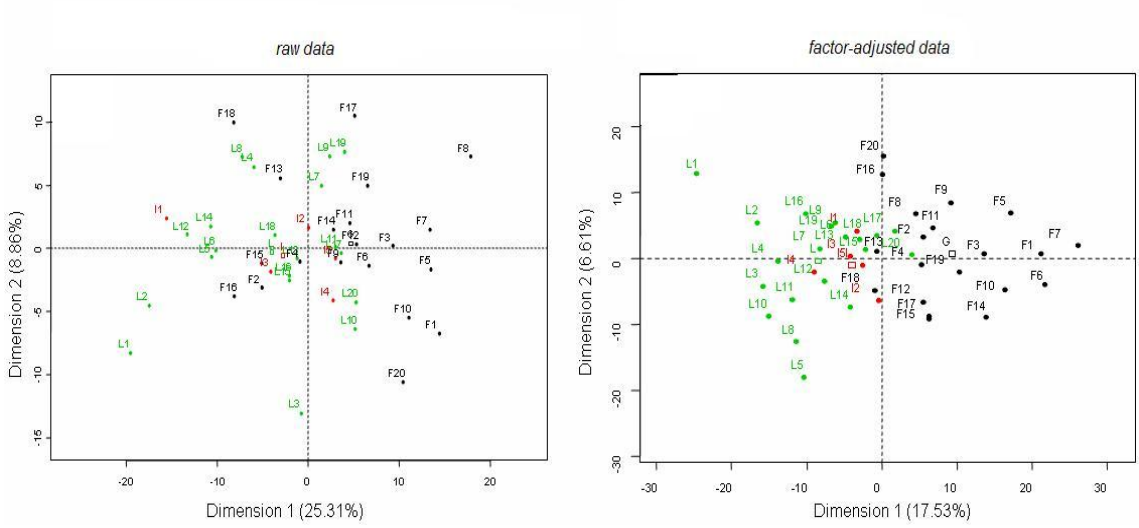


FIGURE 8. *Factorial maps based on gene expressions selected from the raw data (left) and the factor adjusted data (right).*

A more relevant list of genes declared as differentially expressed is of great interest for geneticists

 ISSN: 2102-6238

who carry further analyses of biological properties from this result of differential analysis.

6. Discussion

The issue of this article deals with studying the impact of dependence in large-scale multiple testing procedures. Motivated by issues raised by the analysis of gene expressions data, the aim is to propose a statistical tool to take into account data heterogeneity in simultaneous hypotheses testing for high-dimensional data. Heterogeneity may arise from technical or environmental factors that have not been observed or have not been controlled by the experimental design. The proposed method consists in identifying the linear space generated by a set of latent variables that models the heterogeneous structure, catching the common variability shared by all the response variables. The suggested model is related to a Factor Analysis model.
Most of the existing multiple testing procedures rely on the analysis of the empirical process of p-values associated to the individual tests, under the assumption of independence. A thresholding rule takes into account the multiplicity of the tests. The impact of dependence on the stability of these procedures is one of the prime results of this work. Indeed, dependence induces variability that interferes in particular with p-values distribution under the true null hypothesis. The main impact is a sharp deviation from the theoretical null distribution when the level of common variability between variables is high. Consequently, the variability of false-positives increases. More precisely, the variance of the number of false-positives includes a term which explicitly depends on the correlation between the response variables. Dependence has therefore repercussion on the estimation of error rates, leading to high instability in multiple testing procedures.
A procedure is defined from the factor adjusted variables as the data are independent conditionally on the latent structure. Dependence is actually addressed at the level of the original data, integrated in the model used to calculate the tests statistics. Consequently, the present method illustrates the fact that the well-known individual optimality of the tests indebted to the Neyman-Pearson theory does not imply the global optimality of a multiple testing procedure in situations of dependence between the variables. As data are independent conditionally on the factors, this framework allows to extend to general dependence the results on error rates control initially gained under independence. Thus, the proposed framework leads to less correlation among tests and shows large improvements of power and stability of simultaneous inference. The impact of dependence on the FWER and on π_0 estimation have also been studied and improvements are also brought by the factor analysis framework. These studies are detailed respectively in [11] and in [17].
The proposed procedure is called Factor Analysis for Multiple Testing (FAMT), from the name of the R package. This package is available on the R-project website (`http://cran.r-project.org/`). The package has also its own website (`http://famt.free.fr`). The different steps of the procedures are summed up hereafter:

1. **Estimation of $\mathscr{M}_0$** Classical t-tests are calculated for each variables and a first estimation of $\mathscr{M}_0$ is deduced by taking the indices of the p-values exceeding 0.05;
2. **Choice of the number of factors** The number of factors is estimated by minimization of the criterion given in Proposition 6;
3. **Estimation of the model's parameters** ML estimates are computed considering the EMFA algorithm

 ISSN: 2102-6238

4. **Calculating factor-adjusted p-values** The factor-adjusted test statistics $\tilde{T}$ and the corresponding p-values are calculated;
5. **Up-dating the estimation of $\mathscr{M}_0$** The estimation of $\mathscr{M}_0$ is updated by taking the indices of the factor-adjusted p-values exceeding $0,05$.
   STEPS 2 to 4 are performed again with this new estimation of $\mathscr{M}_0$;
6. **π_0 estimation** The estimation of this parameter is in general a crucial step of multiple testing procedures. Several methods are available (see [24] for details). The estimation is based on the factor-adjusted p-values.
7. **Decision rule** A BH thresholding procedure at level α is applied to the factor-adjusted p-values to decide which null hypotheses are rejected. The BH procedure is improved by plugging-in π_0 estimate.

Beyond the study of the model itself, many points concerning the effective implementation of the factor-adjusted multiple testing procedure are addressed, including the model parameters estimation thanks to an EM algorithm, the estimation of the proportion of true null hypotheses and the choice of the number of factors. The EMFA algorithm provides accurate estimates of variance parameters in a high-dimensional setting (not asymptotic). We propose a criterion allowing to define the model that fits best the covariance structure, minimizing the inflation of variance of false-positives. However, some issues are still to be explored, such as the preliminary estimation of $\mathscr{M}_0$ involved in the calculation of the scores (step 1 in FAMT procedure). This last issue is probably very similar to problems encountered by [15] and by [25].
Finally, the method has been applied to microarray data and great improvements for biological interpretation of differential analysis in gene expressions data have been highlighted [7].

References

[1] D.B. Allison. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39:1–20, 2002.
[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.
[3] Y. Benjamini, A. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93:491–507, 2006.
[4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.
[5] M. A. Black. A note on the adaptative control of false discovery rates. *Journal of the Royal Statistical Society. Series B*, 66:297–304, 2004.
[6] G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic journal of Statistics*, 2:963–992, 2008.
[7] Y. Blum, G. LeMignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11:368, 2010.
[8] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore si Scienze Economiche e Comerciali di Firenze*, pages 3–62, 1936.
[9] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioural Research*, 1:245–276, 1966.
[10] D. Causeur, C. Friguet, M. Houée, and M. Kloareg. Factor analysis for multiple testing (famt): an r package for large-scale significance testing under dependence. *Journal of Statistical Software*, 40(14):1–19, 2011.
[11] D. Causeur, M. Kloareg, and C Friguet. Control of the FWER in multiple testing under dependence. *Communications in Statistics - Theory and Methods*, 38:2733–2747, 2009.
[12] S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

 ISSN: 2102-6238

[13] S. Dudoit, J. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.
[14] B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104, 2004.
[15] B. Efron. Correlation and large-scale simultaneous testing. *Journal of the American Statistical Association*, 102:93–103, 2007.
[16] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.
[17] C. Friguet and D. Causeur. Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Computational Statistics and Data Analysis*, 55:2665–2676, 2011.
[18] C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104:488:1406–1415, 2009.
[19] C. Genovese and L. Wasserman. Operarting characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series B*, 64:499–517, 2002.
[20] I. Hedenfalk, D. Duggan, Y. D. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344:539–548, 2001.
[21] K. I. Kim and M. Van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9, 2008.
[22] E.L. Korn, J.F. Troendle, L.M. McShane, and R. Simon. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124:379–398, 2004.
[23] R. Kustra, R. Shioda, and M. Zhu. A factor analysis model for functional genomics. *BMC Bioinformatics*, 7, 2006.
[24] M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society. Series B*, 67:555–572, 2005.
[25] J. T. Leek and J. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105:18718–18723, 2008.
[26] G. LeMignon, C. Désert, F. Pite, S. Leroux, O. Demeure, G. Guernec, B. Abasht, M. Douaire, P. LeRoy, and S. Lagarrigue. Using transcriptome profiling to characterize qtl regions on chicken chromosome 5. *BMC Genomics*, pages 10–575, 2009.
[27] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. 1979.
[28] R. G. Montanelli and L. G. Humphrey. Latent roots of ranrom data correlatoin matrices with squared multiple correlations on the diagonal: a monte-carlo study. *Psychometrica*, 41:341–348, 1976.
[29] A.B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society. Series B*, 67:411–426, 2005.
[30] K. Pollard, Y. Ge, S. Taylor, and S. Dudoit. *multtest: Resampling-based multiple hypothesis testing*. R package version 1.23.3.
[31] I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8:61, 2007.
[32] D. Robertson and J. Symons. Maximum likelyhood factor analysis with rank-deficient sample covariance matrix. *Journal of Multivariate Analysis*, 98:813–828, 2007.
[33] D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47:69–76, 1982.
[34] J. Shaffer. Multiple hypotheses testing: a review. *Annual review of psychology*, 46:561–584, 1995.
[35] C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.
[36] J. Storey and R Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.
[37] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64:479–498, 2002.
[38] J. D. Storey. The positive false discovery rate: a bayesian interpretation and the q -value. *Annals of Statistics*, 31:2013–2035, 2003.
[39] J. D. Storey, J.Y. Dai, and J. T. Leek. The optimal discovery procedure for large-scale significance testing, with application to comparative microarray experiments. *Biostatistics*, 8:414–432, 2007.
[40] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous

 ISSN: 2102-6238

conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B*, 66:187–205, 2004.

[41] J.D. Storey. The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society. Series B*, 69:347–368, 2007.

 ISSN: 2102-6238